IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Issue-1 E-ISSN: 2347-2693

Classifying the Incidence Rates of Cancers Using Data Mining Techniques (Perspective to Gas Leakage Accident of Bhopal City)

Sanjeev Gour

Department of Computer Science, Career College, Bhopal, India

Corresponding Author: Sunj129@gmail.com

Available online at: www.ijcseonline.org

Received: 14/Dec/2017, Revised: 25/Dec/2017, Accepted: 15/Jan/2018, Published: 31/Jan/2018

Abstract-Cancer is one of most dangerous diseases and the incidence rate of cancer in India is increasing every year. Today, number of tools and techniques are available to analyze large cancer dataset. Data mining tools are most frequently used to identify patterns in cancer patients and in cancer diagnosis or detection. Data mining is now widely used in health care industry as it has a great capability to extract hidden patterns from the large past medical record of cancer patients. This study uses Data Mining techniques called Classification and Clustering to classify and compare the incidence rates of TCR (Tobacco Related Cancer) and Non-TCR (Non-Tobacco Related Cancer) in two areas of Bhopal city and extracted some useful and interesting fact from the incidence rates data of cancer patients. The incidence rate data of cancer were obtained during past 40 years from Bhopal-population-based cancer registries (PBCR) of two areas partitioned after Gas tragedy of Bhopal city. This study can be helpful to the medical analysts as a decision support system. The study was done using WEKA Tool.

Keywords: *Data Mining, Classification, Clustering, WEKA, Tobacco Related Cancer.*

I. INTRODUCTION

In medical Science, the information of medical related records are extremely rich and sound but there are only few powerful tools to identify hidden patterns and relationship in medical data. Data mining methods have a large scope in the field of disease diagnosis and health care patterns. This study proposed the application of Data mining in healthcare sector. The most frequently used data mining method is classifications that classified or predict the categorical class. In this study attribute called 'Area ***(MIC affected)*** 'is the class label. The other Technique is clustering which is used to identify similar objects in one group called cluster while non-similar objects in another cluster within the data set [1].In this study, there are two clusters were generated for two areas respectively. One cluster is for MIC affected area and other is for Non-MIC affected area.

Cancer is one of the most common and dangerous diseases in the world that causes majority of death. In many research studies, it is shown that a cancer incidence rate varies between people on the basis of different factors. These can be socioeconomic level, educational level or genetically. It can also by any chemical industry disaster like Bhopal Gas leakage tragedy which is world's most industrial disaster.

According to Free press journal (4th Feb. 2016) Bhopal is at second position in the country as 74.2% cases were belong to Tobacco Related Cancers (TRC).So analysis of the incidence data of Tobacco related cancer can be very helpful in policy making in the field of Medical science [2]. For this purpose, Data mining techniques have great capability to uncover patterns hidden in the data and to identify the relationship among various parameters of data and this can help the medical analyst in decision making. From study [3] it is detected that the correctness for the cancer diagnosis and analysis of cancer patterns of various applied data mining classification methods is favorably acceptable and this can help the medical analyst in decision making for making systematic medical policies.

In 1984, a chemical industrial disaster in the factory of union Carbide in Bhopal causes total of 3,787 deaths related to the gas release in the city. Methyl IsoCtanate (MIC) was assumed to be the main content of the toxic gas. After this accident, the Bhopal municipal area was partitioned into two areas called MIC affected (Area-1/Area-E) and MIC unaffected (Area-2/ Area-U) area .After this industrial disaster, many research agencies initiated studies to evaluate the effect of this accident. Along with the same intension, the incidence rates of TRC and Non- TRC classified with respect to two areas (Area-E and Area-U) of Bhopal city using the classification and clustering algorithms of Data Mining.

II. DATA MINING IN MEDICAL SCIENCE

Today Healthcare industry creates huge volumes of complex data includes electronic patient records, clinical resources, diagnosis-analysis data, medical devices etc. and this High volume of data need to be analyzed for extraction of useful information which used or helpful in decision support system of [4].

Although from more than 40 years, many data mining techniques and tools have been applied in many fields and sectors already but their applications in medical science are comparatively new. In the last decade, in healthcare or medical data mining, massive growth of application of data mining took place [5].

Data Mining is now one of the most leading technology in research areas [6] and that is also became most common in medical science. The DM techniques play significant role for identifying hidden patterns and relationship among different medical parameters of large medical dataset of any healthcare organization so that these mined outcomes are very useful to carry out efficient medical diagnosis.

The analysis of medical processes is completely based on the identification of hidden patterns and relationship which are present in the data. So for diagnostic procedure or analysis any medical process, these patterns may be very useful. As data mining include many pattern recognition methods, it is the best technology today to mine the medical dataset [7].

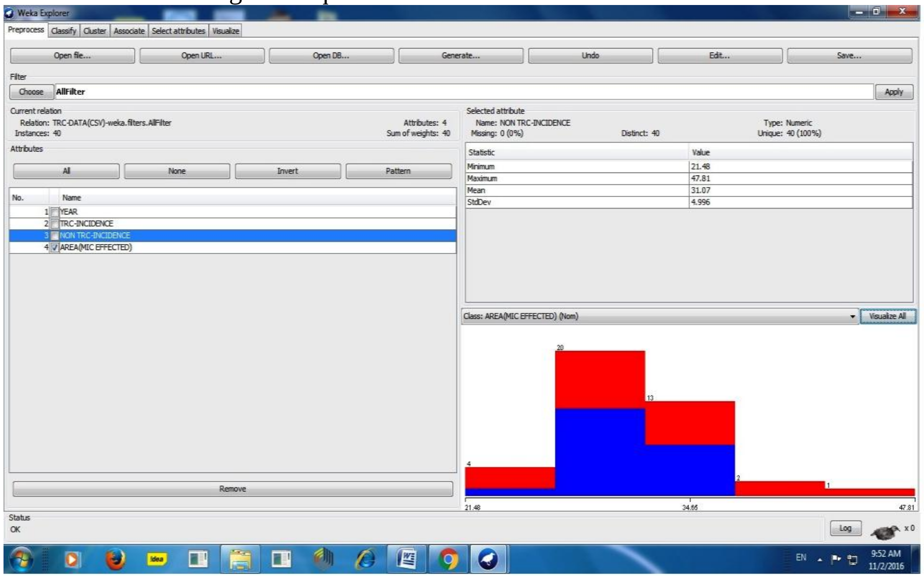
Some of applications of Data mining in Medical science are discussed here: P. Ramachandran et. al. [8] proposed and builds a cancer risk prediction model by using classification and clustering methods of DM. Their model proposed a cost effective and easy way for diagnosis procedure for different types of cancer and also gave a effective preventive blueprint for any healthcare process. Joseph A. et.al [9] reviewed the procedure of cancer prediction by explaining, comparing and assessing the performance of different DM techniques applied to cancer prediction. Anupama et.al [10] found in their study that there is a broad scope to carry out research work in the field of cancer diagnosis using data mining techniques. Moh.Shabaz [11] uses data mining tool and technique to build a decision support system to recover and locate different types of cancer patterns on the Genes dataset. In this study, author also used data mining techniques named classification and clustering which help in classifying and comparing the incidence rates of cancer in two areas of city.

III. DATASET

The data for this study were obtained from published report of Bhopal population-based cancer registries (PBCR) in excel format for past 40 years. After the industrial chemical tragedy in which MIC gas has been open out into the atmosphere at Bhopal city, a PBCR was started under the National Cancer Registry Programme reports. Incidence rate of cancer indicates new cases diagnosed or registered in a specified population in a defined time period. For this study all cancer cases registered from 1988 to 2007 in the Bhopal PBCR, are included. The incidence rates of cancer usually indicated as Age Adjusted Rates (AAR) per 100000 persons and these rates can be calculated by taking the age specific rates and applying these rates to the standard population belongs to specific age group. The incidence rates of cancer data of Bhopal PBCR was then categorized into two named Area-E and Area-U subjected to MIC affected and Non-affected area respectively of Bhopal city. For this study, four attribute were taken (Table-1).

Table 1: Sample format of cancer Dataset

Year	TRC-Incidence rate	Non-TRC Incidence rate	Area(MIC affected)
1988	24.05	27.44	Area-E /Area-1
-----	-----	------	------
1995	28.02	32.97	Area-U/Area-2
-----	-----	------	------
2007	30.55	47.81	Area-U/Area-2



Filter: AllFilter

Current relation
Relation: TRC-DATA(CSV)-weka.filters.AllFilter Attributes: 4
Instances: 40 Sum of weights: 40

Attributes

No.		Name
1	□	YEAR
2	□	TRC-INCIDENCE
3	□	NON TRC-INCIDENCE
4	☑	AREA(MIC EFFECTED)

Selected attribute
Name: NON TRC-INCIDENCE Type: Numeric
Missing: 0 (0%) Distinct: 40 Unique: 40 (100%)

Statistic	Value
Minimum	21.48
Maximum	47.81
Mean	31.07
StdDev	4.996

Class: AREA(MIC EFFECTED) (Nom)

Status
OK

Figure 1. Parameter/Attribute view in WEKA

IV. ABOUT DATA MINING AND ITS TOOL

Conceptually data mining is focused with the analysis of data and the use of software methods or algorithms for finding unknown patterns and features in large sets of data.

Data mining examine the data from different views and concludes it into piece of interesting and useful information. The methods of machine learning are similar to methods of data mining but data mining methods obtain the data for human conception.

WEKA is one of data mining software tool which is a collection of machine learning methods. These methods can be directly applied to the user data and extract some or many hidden patterns from the large dataset. It is most frequently used tool nowadays because a beginner can go through the process of applied machine learning using the graphical interface without having to do any programming. For this study, author used the WEKA 3.8.1 window based (64 bit) version.

To classify the cancer data set based on the characteristics of attributes, author used two data mining methods called Classification and Clustering thought WEKA tool.

Classification: To predict or classify nominal or numeric values from given dataset. The classification algorithms called Random Tree, J48 used for this study.

Clustering: This method identifies similarities or groups of occurrences within the data set. Clustering algorithms called ***SimpleKmean*** and Canopy used for this study.

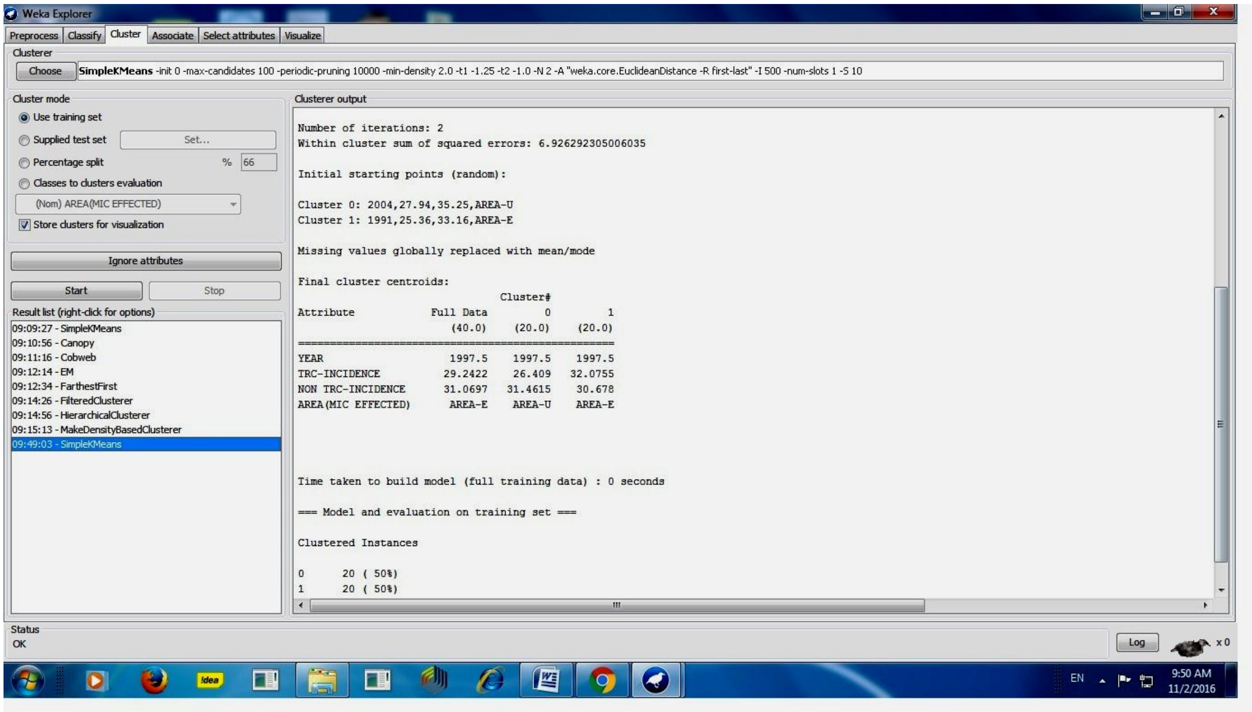
V. EXPERIMENTAL PROCESS

The data set used for this study is processed in comma-separated format (CSV). The resulting data file (TRC-DATA.csv) includes 40 instances of TRC and Non-TRC incidence rates for 40 Years and these are categorized in two areas named Area-1 and Area-2 based on MIC affected and Non-affected respectively (Table-1).

In WEKA Explorer environment, first the training data file named "TRC-DATA.CSV" loaded and clean and assign the attribute as Class label by ***Preprocess*** mode. After that author selects a one of clustering technique, named ***SimpleKMeans,*** and by pressing the ***Start*** button we got the clustering result in its output window. The same procedure was done for Canopy method. In this experiment, WEKA classifies the training data or samples into clusters according to the cluster parameter setting and calculates the percentage of instances falling in each cluster. In this experiment ,results produced in output window by k-means shows 50% (20 instances) in cluster-0 and 50% (20 instances) in cluster-1 (figure-2).It assigns instances of given trained dataset to the clusters, based on the similarities value of the that class attribute within each cluster. Canopy clustering generates three clusters with instance of 35%, 50% and 15 % respectively (figure-3).

For classification the cancer dataset first, load the training data file and assign the Class attribute as Class label in WEKA Explorer environment. Select the **Classify** mode, and choose the classification algorithm, named Random Tree.

Now it is need to implement the classification by clicking the ***Start*** button and we got the classification result in its output window (figure-4). The size of resultant tree is 19. Correctly classified instance are 100 %. The random trees classifier receive the trained instances of data, classifies it with every tree in the forest, and gives the class label as output that accept the majority values from the dataset. This method clearly classifies the incidence rates of cancer of TRC and non-TRC within two areas of city with respect to time in years (from 1988 to 2007).The same procedure was done for J48 algorithm of data mining. In this case, the Tree size is 9 and correctly classified instances are 95 %. This is a predictive algorithm which classifies the target sample to a new sample based group (Class Label) on different attribute values of the available data.



Clusterer

SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Cluster mode

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Nom) AREA(MIC EFFECTED)
- ☑ Store clusters for visualization

Result list (right-click for options)

```
09:09:27 - SimpleKMeans
09:10:56 - Canopy
09:11:16 - Cobweb
09:12:14 - EM
09:12:34 - FarthestFirst
09:14:26 - FilteredClusterer
09:14:56 - HierarchicalClusterer
09:15:13 - MakeDensityBasedClusterer
09:49:03 - SimpleKMeans
```

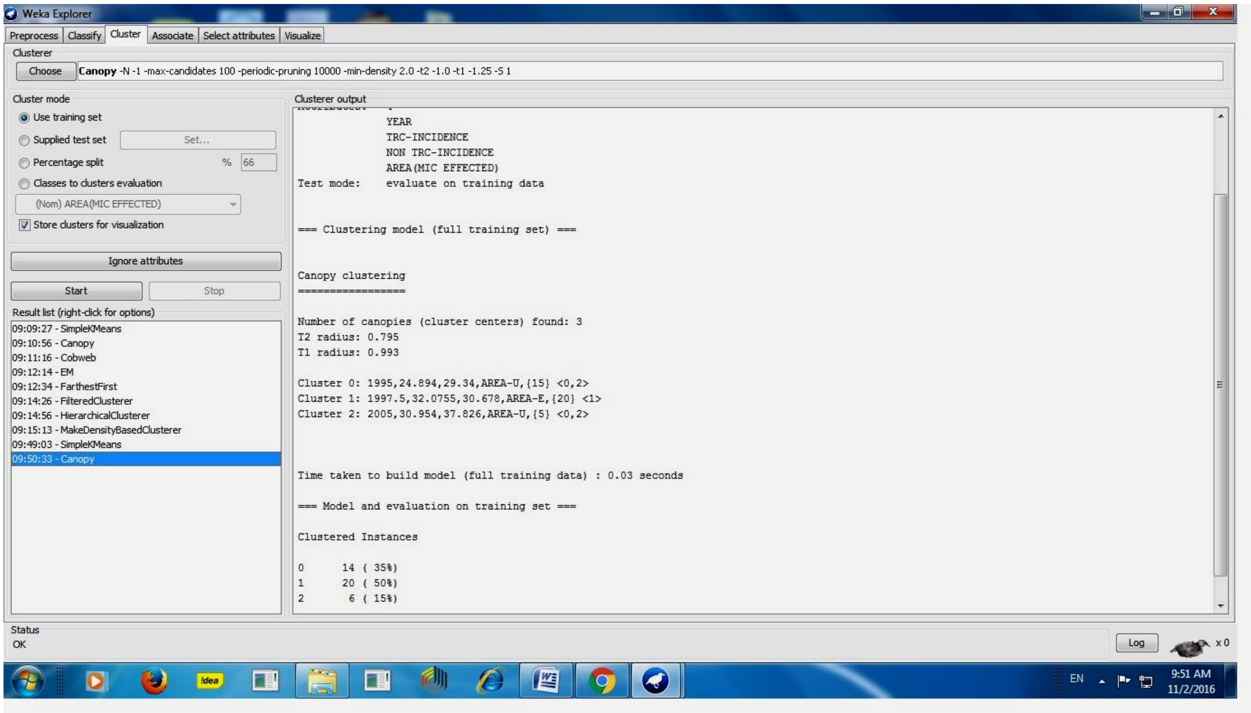
Clusterer output

```
Number of iterations: 2
Within cluster sum of squared errors: 6.926292305006035

Initial starting points (random):

Cluster 0: 2004,27.94,35.25,AREA-U
Cluster 1: 1991,25.36,33.16,AREA-E

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute           Full Data          0          1
                       (40.0)     (20.0)     (20.0)
=====================================================
YEAR                   1997.5     1997.5     1997.5
TRC-INCIDENCE         29.2422     26.409    32.0755
NON TRC-INCIDENCE     31.0697    31.4615     30.678
AREA(MIC EFFECTED)     AREA-E     AREA-U     AREA-E




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      20 ( 50%)
1      20 ( 50%)
```

Status

OK

Figure 2. Implementation View in WEKA for clustering (k-means)



Clusterer

Canopy -N -1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1

Cluster mode

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Nom) AREA(MIC EFFECTED)
- ☑ Store clusters for visualization

Result list (right-click for options)

```
09:09:27 - SimpleKMeans
09:10:56 - Canopy
09:11:16 - Cobweb
09:12:14 - EM
09:12:34 - FarthestFirst
09:14:26 - FilteredClusterer
09:14:56 - HierarchicalClusterer
09:15:13 - MakeDensityBasedClusterer
09:49:03 - SimpleKMeans
09:50:33 - Canopy
```

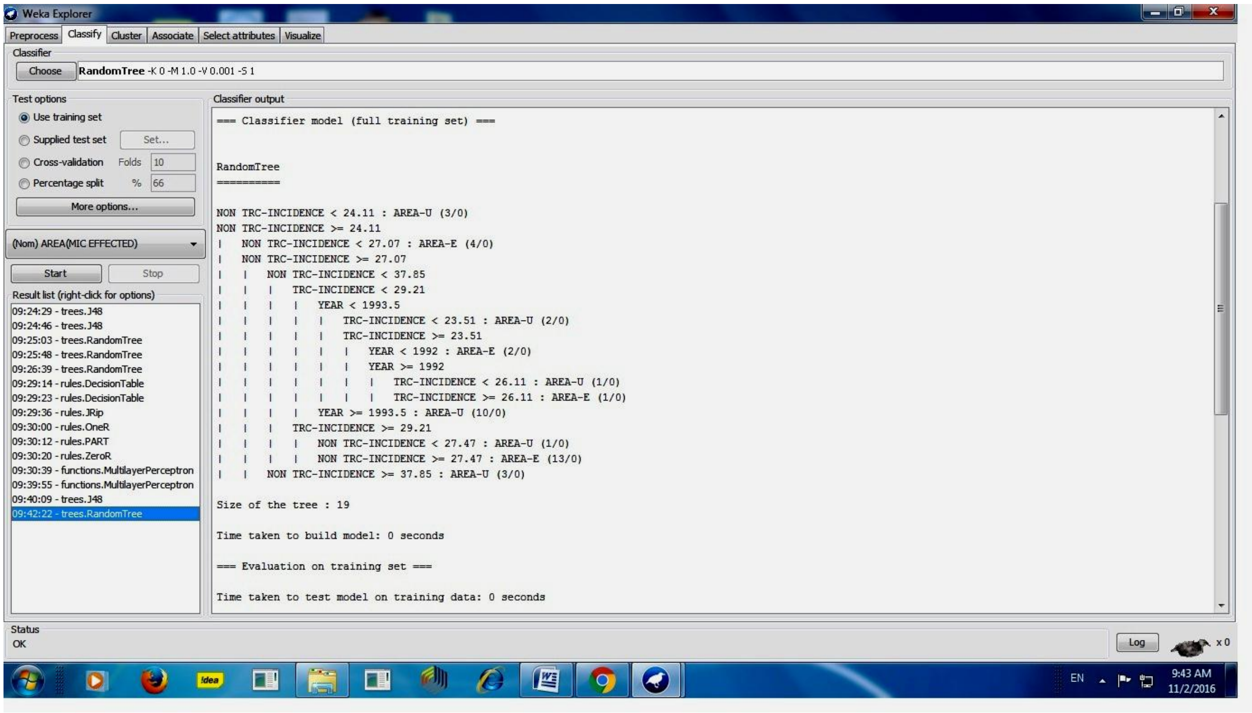
Clusterer output

```
              YEAR
              TRC-INCIDENCE
              NON TRC-INCIDENCE
              AREA(MIC EFFECTED)
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


Canopy clustering
=================

Number of canopies (cluster centers) found: 3
T2 radius: 0.795
T1 radius: 0.993

Cluster 0: 1995,24.894,29.34,AREA-U,{15} <0,2>
Cluster 1: 1997.5,32.0755,30.678,AREA-E,{20} <1>
Cluster 2: 2005,30.954,37.826,AREA-U,{5} <0,2>



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      14 ( 35%)
1      20 ( 50%)
2       6 ( 15%)
```

Status

OK

Figure 3. Implementation View in WEKA for clustering (Canopy)



Classifier: RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Test options
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) AREA(MIC EFFECTED)

Result list (right-click for options)

```
09:24:29 - trees.J48
09:24:46 - trees.J48
09:25:03 - trees.RandomTree
09:25:48 - trees.RandomTree
09:26:39 - trees.RandomTree
09:29:14 - rules.DecisionTable
09:29:23 - rules.DecisionTable
09:29:36 - rules.JRip
09:30:00 - rules.OneR
09:30:12 - rules.PART
09:30:20 - rules.ZeroR
09:30:39 - functions.MultilayerPerceptron
09:39:55 - functions.MultilayerPerceptron
09:40:09 - trees.J48
09:42:22 - trees.RandomTree
```

Classifier output

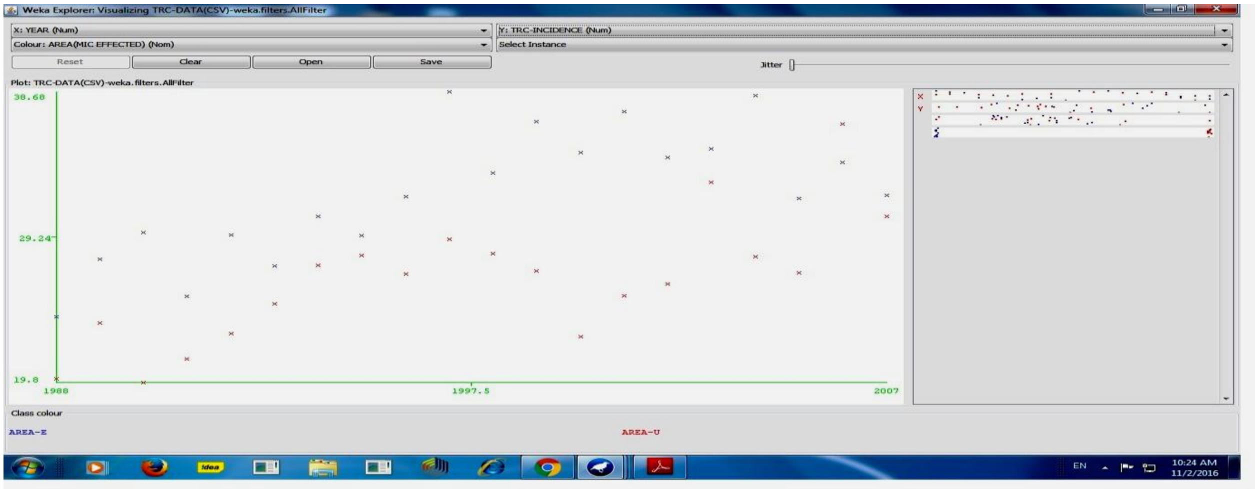
```
=== Classifier model (full training set) ===


RandomTree
==========

NON TRC-INCIDENCE < 24.11 : AREA-U (3/0)
NON TRC-INCIDENCE >= 24.11
|   NON TRC-INCIDENCE < 27.07 : AREA-E (4/0)
|   NON TRC-INCIDENCE >= 27.07
|   |   NON TRC-INCIDENCE < 37.85
|   |   |   TRC-INCIDENCE < 29.21
|   |   |   |   YEAR < 1993.5
|   |   |   |   |   TRC-INCIDENCE < 23.51 : AREA-U (2/0)
|   |   |   |   |   TRC-INCIDENCE >= 23.51
|   |   |   |   |   |   YEAR < 1992 : AREA-E (2/0)
|   |   |   |   |   |   YEAR >= 1992
|   |   |   |   |   |   |   TRC-INCIDENCE < 26.11 : AREA-U (1/0)
|   |   |   |   |   |   |   TRC-INCIDENCE >= 26.11 : AREA-E (1/0)
|   |   |   |   YEAR >= 1993.5 : AREA-U (10/0)
|   |   |   TRC-INCIDENCE >= 29.21
|   |   |   |   NON TRC-INCIDENCE < 27.47 : AREA-U (1/0)
|   |   |   |   NON TRC-INCIDENCE >= 27.47 : AREA-E (13/0)
|   |   NON TRC-INCIDENCE >= 37.85 : AREA-U (3/0)

Size of the tree : 19

Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds
```

Status
OK

Figure 4. Implementation View in WEKA for classification

VI. RESULT AND DISCUSSION

From the results of these algorithms implemented on WEKA tool, It is clearly observed that in 1991 NON-TRC incidence rate is higher than TRC incidence in affected Area 1 (AREA-E) while all TRCs over the years observed a significantly higher in Area 1 (AREA-E) as compared to Area 2(AREA-U) while it is gradually decreases as compared to TRC rates during and after year 1995. This result clearly state that gas leakage incident is obviously one of major reason to increase the Non-TRC incidence rates in years after 1984. The results from (figure-4) random tree classification also shows that the TRC rates is significantly increase in 10 years from 1995 to 2005 as compared to Non-TRC in both areas as the results of survey indicates that approximate. 50 % of male (above 21 years of age) in AREA-E and 34 % of males of same age group in AREA-U were tobacco users. Form the figure-5, X-Y plot for class attribute "Area *(MIC –affected)"*, it is clearly shown that the incidence rates of non-TRC cancer is higher in Area-E during the year from 1988 to 1997 while the population consuming tobacco is higher in both areas. And after year 1997 it is showing also from this plot that incidence rates of TRC is as proportion to the population consuming tobacco is approximately in similar with rates of Non-TRC.



X: YEAR (Num)
Y: TRC-INCIDENCE (Num)
Colour: AREA(MIC EFFECTED) (Nom)

Plot: TRC-DATA(CSV)-weka.filters.AllFilter

Class colour
AREA-E AREA-U

Figure 5. X-Y plot for Class attribute ***Area*** with ***Year*** as-axis and ***Incidence rate*** as X-axis.

VII. CONCLUSIONS

In this paper data mining techniques like classification and clustering were used to classify the incidence rates of Tobacco related cancer (TCR) and Non-Tobacco related cancer(Non-TCR) in two areas of Bhopal city and conclude from the study that incidence rates of Tobacco related cancer are higher than in the affected area 1as compared to unaffected area 2 while in during the periods form 1984 to 1994 (10 year after Gas leakage incident) NON-TRC incidence rate is higher than TRC incidence in affected Area 1 (AREA-E) as compared to unaffected Area 2 .so this study conclude that gas leakage incident is obviously one of major reason to increase the Non-TRC incidence rates in years after 1984 in affected Area 1.However, when the Tobacco habits of the populations in the two areas were taken into account, the higher rates observed in the MIC affected area was essentially neutralized. This result indicate that the higher AAR observed in the MIC affected area was due to the higher proportion of the population consuming tobacco rather than due to the effect of the MIC exposure. Also as the majority of the populations of the affected area are from a lower socioeconomic group compared to the unaffected area. In this way this study may also be helpful for medical analyst to focus and concentrate more on medical- history records of TCR and Non-TCR patients after Gas Leakage accident of Bhopal city.

REFERENCES

[1] N. Kumar, S. Khatri, "Implementing WEKA for medical data classification and early disease prediction-computational intelligence and communication Technology, 10 feb-2017.

[2] S. Asthana, R.S.Patil, S.Labani, " Tobacco‑related cancers in India: A review of incidence reported from population‑based cancerregistries", Indian Journal of Medical and Paediatric Oncology. Volume 37, Issue 3, pp-152-157, Jul-Sep 2016.

[3] S.Gupta, D.Kumar, A.Sharma, "Data Mining Classification Techniques Applied For Breast Cancer Diagnosis And Prognosis**",** Indian Journal of Computer Science and Engineering **(IJCSE)** Vol. 2 No. 2,pp.-188-195, Apr-May 2011.

[4] M. Durairaj, V. Ranjani,." Data Mining Applications In Healthcare Sector: A Study", International Journal Of Scientific & Technology Research ,Volume 2, Issue 10, pp.-29-35, October 2013.

[5] O.Niakšu, O. Kurasova, " Data Mining Applications in Healthcare: Research vs Practice", Data Mining Application in Healthcare Research vs Practice, pp.-58-70, 2014.

[6] S. Gour , "*Developing Decision Model by Mining Historical Prices Data of Infosys for Stock Market Prediction*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.10, pp.92-97, 2016.

[7] S. L. Ting, C. C. Shum, S. K. Kwok, A. H. C. Tsang, W. B. Lee, " Data Mining in Biomedicine: Current Applications and Further Directions for Research- scientific Research", J. Software Engineering & Applications, 2, pp.- 150-159, 2009.

[8] P.Ramachandran, N. Girja, T.Bhuvaneswari, "Early Detection and Prevention of Cancer using Data Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 97– No.13, pp.-48-53, July 2014.

[9] A. Joseph, S. David," Applications of Machine Learning in Cancer Prediction and Prognosis", Departments of Biological Science and Computing Science, University of Alberta Edmonton, AB, Canada, Cancer Informatics, pp.- 59-77, 2006.

[10] Y.K. Anupama, S.Amutha,R.Babu ," Survey on Data Mining Techniques for Diagnosis and Prognosis of Breast Cancer" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 5, Issu-2, pp.- 33 – 37, Feb-2017.

[11] M. Shabaz, S.Faruq,M.Shaeen, S.A. Masood, "Cancer Diagnosis Using Data Mining Technology" Life Science Journal, vol-9, issue-1, pp-308-313, September 2012.

Author Profile

Sanjeev Gour pursed Bachelor of Computer Science (Honors) and Master of Science (Electronics & Comm.) from Devi Ahilya University Indore and Master of Science (Computer Science) from Barkatullah University Bhopal. He has completed his Ph.D. in Computer Science and currently working as a Asst.Professor in Department of Computer Science in Career College, Bhopal. He was member of Board of Studies & Examination Committee of Computer Science in BU Bhopal University also a member of Managing Committee in Computer Society of India (Bhopal Chapter). He has published more than 13 research papers in reputed international journals including Thomson Reuters & Scopus (SCI & Web of Science).His main research work focuses on Data Mining. He has 15 years of teaching experience and 4 years research experience.